DOCUMENT RESUME

ED 376 203 TM 022 312

AUTHOR Hamilton, Laura S.
TITLE An Investigation of Students' Affective Responses to Alternative Assessment Formats.
SPONS AGENCY National Science Foundation, Washington, D.C.; Rand Corp., Santa Monica, CA. Inst. for Education and Training.
PUB DATE Apr 94
CONTRACT MDR-9154406; RED-9253068
NOTE 28p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 5-7, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Affective Behavior; Constructed Response; Educational Assessment; *Elementary School Students; *Emotional Response; High Schools; High School Students; Intermediate Grades; Interviews; Mathematics; Multiple Choice Tests; State Programs; *Student Attitudes; *Test Format; Testing Programs
IDENTIFIERS Alternative Assessment; Hands on Science; Performance Based Evaluation; Student Engagement

ABSTRACT

Despite the number of studies investigating affective aspects of test taking, little is known about how students perceive the kinds of extended performance assessments currently being developed for state and local testing programs. This paper presents two studies that address these issues. In the first, hands-on science tasks were administered to 20 sixth-grade and 29 fifth-grade students who thought aloud as they performed each task and answered interview questions afterward. In the other study, mathematics items were administered in three formats (multiple choice, short-answer constructed response, and extended problems) to 29 high school students who were interviewed after completing the items in each format. Results of both studies indicate a great deal of variability in the affective responses of students to novel assessment formats, and they suggest some possible influences on these responses, including the importance of the nature of engagement and students' perceptions of validity and fairness. Three tables and one figure present study findings. (Contains 16 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 376 203

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

LAURA S. HAMILTON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

An Investigation of Students' Affective Responses to Alternative Assessment Formats

Laura S. Hamilton

Stanford University

Paper presented at the 1994 annual meeting of the National Council on Measurement in Education, New Orleans.

This research was supported in part by National Science Foundation grants MDR-9154406 and RED-9253068, and by the RAND Institute on Education and Training. The science tasks used in this study were developed by researchers at RAND and at the University of California, Santa Barbara. I am grateful to Dr. Edward Haertel and Dr. Richard Snow for their comments on an earlier draft and to Michele Ennis, Haggai Kupermintz, and Andrea Buell for their assistance in data collection.

TM022312

BEST COPY AVAILABLE

An Investigation of Students' Affective Responses to Alternative Assessment Formats

Traditional paper-and-pencil tests are viewed by many critics as inadequate measures of students' abilities because they fail to elicit complex thinking and deep subject-matter understanding (e.g., Frederiksen, 1984; Resnick & Resnick, 1992). Traditional tests have also been criticized for their failure to capture students' interest or to involve them in activities that are intrinsically motivating (Gardner, 1992). Interest and motivation have been shown to affect test performance, and therefore we might expect tasks that are more engaging to elicit better performance. In addition, many reformers believe that an emphasis on performance-based assessment will narrow performance gaps among students of different ethnicities. The achievement of this goal is at least partially contingent on increasing the motivation of traditionally low-achieving groups to perform well on new tests. This paper reports some preliminary findings regarding student attitudes toward different assessment formats, and explores the relationships between these attitudes and performance.

Previous studies have uncovered important ways in which affective and motivational variables affect test performance. Studies of motivational variables such as test anxiety and self-efficacy have been reviewed by Crooks (1988) and by Snow (1989). Crooks, for example, describes curvilinear relationships between some motivational variables and achievement. Of particular relevance to current testing reform efforts are findings that format differences may influence students' experiences anxiety or other potential obstacles to performance. Moreover, these reactions may differ according to experience or background. In a study of the 1990 NAEP Mathematics assessment, Koretz, Lewis, Skewes-Cox, and Burstein (1992) found that members of some ethnicities were less likely to respond to open-ended items than were students in other groups. This finding suggests that the experiences students bring to the testing situation may interact with test format to influence their performance, and that elimination of the multiple-choice format may increase, rather than reduce, achievement gaps.

Many critics assert that students find multiple-choice tests boring and irrelevant to their lives (Gardner, 1992). Certainly the act of filling in circles on a sheet of paper is not one in which we expect students to be engaged after they leave school, and most multiple-choice tests are not designed to be intrinsically interesting. In contrast, Wiggins (1992) includes in his list of criteria for test design meaningfulness to students and presentation of problem contexts that are "rich, realistic, and enticing" (p.27). He rejects the notion that tests must appear to be directly relevant to students' lives, but claims that they should be meaningful and engaging in order to elicit maximum involvement and performance. Linn, Baker, and Dunbar (1991) also urge assessment researchers to consider the meaningfulness of tests to students.

Wiggins stresses the importance of clear performance standards: Students should know what constitutes an acceptable response, without having to wonder whether they are proceeding correctly or providing sufficient detail in their answers. Frederiksen and Collins (1989) refer to this quality as *transparency*. This is essential to consider as new assessment formats are introduced, because although multiple-choice tests do not make the correct answers obvious, most students know what is expected of them when they take such tests. Performance assessment may elicit anxiety not because of any characteristic inherent in the format, but because students lack experience with these tests and awareness of the expectations that are placed on them.

There is also evidence that test format sends an important message to students about how they should prepare. Snyder (1971) describes testing as creating a "hidden curriculum" that informs students about what they are expected to learn. D'Ydewalle, Swerts, and De Corte (1983) found that students who expected a free-response test performed better on both free-response and multiple-choice measures than did those expecting a multiple-choice test. Because this performance difference could not be attributed solely to study time, the authors suggest that expectations affected processing of the study material. Rocklin (1992) found that the primary dimension along which college students distinguish items is supply versus selection, and that they tend to

perceive selection items (e.g., multiple-choice or true-false) as less difficult and more objective than items requiring them to supply a response. He suggests that these views may stem from experience with the formats or from perceptions of the amount of cognitive processing needed to produce responses. Rocklin also states that individual differences in students' perceptions result in different degrees of influence of format on method of studying. Lundeberg and Fox (1991), however, note that test expectancy effects are more salient for laboratory than for classroom tests, and suggest that perceived type of thinking required may be an important predictor of quality of studying that is related but not identical to format. Thus it is important to investigate students' perceptions of the kind of thinking required by different test formats.

In addition, it is essential that we consider students' preferences for different formats, because this may influence their investment in the test and, consequently, their performance. A survey designed and administered by Check (1982) indicated that students prefer multiple-choice tests to other formats, and that extended essays are least preferred. These results may relate to perceived difficulty; students may prefer tests that require less effort or induce less stress. However, preferences may stem from other features of test formats such as perceived fairness or the susceptibility of the multiple-choice format to test-taking strategies. Bridgeman (1992) found that 81% of students who took stem-equivalent multiple-choice and constructed-response items from the Quantitative section of the Graduate Record Examination preferred the multiple-choice, and 88% of this same sample reported using strategies such as working backward. Interestingly, only 43% believed that the multiple-choice format was fairer, and 41% thought that the constructed-response was fairer. Apparently, students do not always prefer the test that they believe is fairest.

Despite the number of studies investigating affective aspects of test-taking, little is known about how students perceive the kinds of extended performance assessments currently being developed for state and local testing programs. This paper presents two studies that address

these issues. In one study, hands-on science tasks were administered to fifth- and sixth-grade students, who thought aloud as they performed each task and answered interview questions afterward. In the other study, mathematics items were administered in three formats to high school students, who were interviewed after completing the items in each format. These studies should be regarded as preliminary efforts to gather students' opinions about assessment tasks, to explore relationships between attitudes and performance, and to suggest areas for future research.

Study 1: Elementary School Hands-On Science Tasks

Method

Instruments:

The tasks used in this study were developed for a larger study of hands-on science assessment, and the interview data reported here were collected as part of this larger study. The tasks were designed to be administered to groups of students in their classrooms; students record their responses in booklets that are later scored. For the present study, however, students took the tasks individually and were asked to "think aloud" and to answer questions about their responses afterward, so that an the cognitive demands of the tasks could be analyzed.

Four 6th-grade tasks were administered, two from each of two task shells. For the purposes of this study, a *task shell* is defined as a set of specifications that describes the format of the task, the skills to be tested, and the criteria for judging the adequacy of responses. The Pendulum and Lever tasks, developed out of an "Inference" shell, require the student to conduct an experiment and to answer questions based on the results. Students are given instructions that permit them to observe the effects of two independent variables on one dependent variable, and must identify which independent variable is the important one. On Pendulum, students construct four pendulums using strings of two different lengths and sets of washers of two different weights, and must determine whether the string length or the weight affects the

amount of time needed for the pendulum to swing 20 times. On Lever, students test the lifting ability of four levers, constructed from bars on which length and fulcrum location are varied. Table 1 shows the inference questions for Pendulum and Lever.

Table 1

Inference Questions

PENDULUM:	LEVER:
1. Which two pendulums took the most time to swing 20 times?	1. Which two levers needed the most washers to lift the weight?
2. Dale says the weight of the pendulum has the biggest effect on how fast it swings. Pat says the length of the string is more important. Who is right? Explain your answer.	2. Chris says the length of a bar has the biggest effect on its ability to lift objects. Jody says the location of the notch is more important. Who is right? Explain your answer.
3. Look at Pendulum E on the cardboard. How much time would it take Pendulum E to swing 20 times?	3. Look at Bar E on the cardboard. How many washers will it take to lift the weight with this bar?

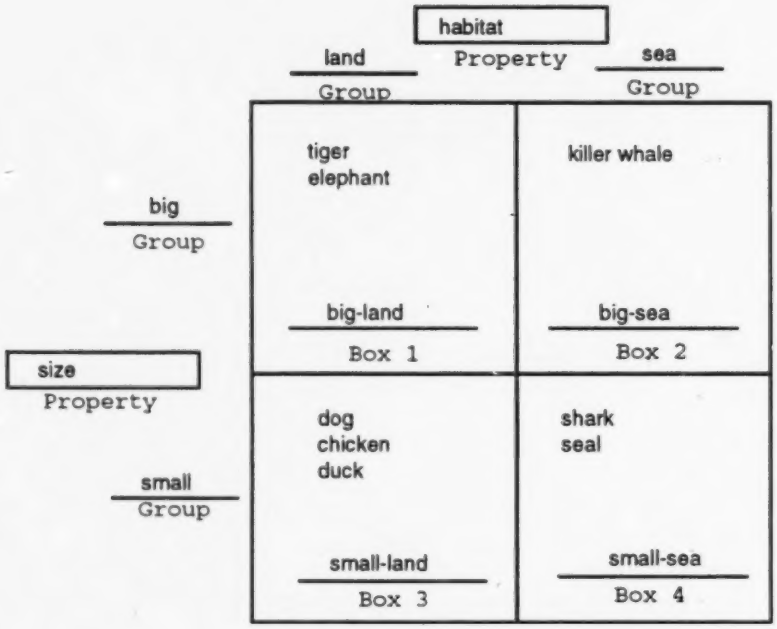
The other 6th-grade tasks, Animals and Materials, were developed from a two-way cross-classification shell. Students are first given a tutorial that provides examples of the cross-classification activity using pictures of people, and asks them to complete some partially-finished cross-classification tables. On the second part of the task, students are given eight animals (e.g., tiger, seal) or eight materials (e.g., pine cone, seaweed) with which they are asked to form a two-way cross-classification. The student chooses the variables by which to classify, and is given a large placemat that he or she may use for sorting the objects. Students must write the names of the objects in a table in the test booklet, and must label the parts of the table, as illustrated in Figure 1. The last question asks the student to remove a

9th object from an envelope and to classify it according to the system that he or she has created.

Two fifth-grade tasks from one shell were administered: Friction and Incline. Each task provides open-ended instructions that ask students to design their own experiments to answer a specific question.

Figure 1

Two-Way Cross-Classification Table



Unlike the 6th-grade Inference tasks, Friction and Incline do not give the students explicit instructions concerning which observations to conduct. Students must decide which variable to focus on and must record the steps in their experiments. They are then given results from a completed experiment, which they use to construct a graph and to answer some interpretation questions.

Procedures

The 6th-grade tasks were administered individually to 20 6th-grade students, and the 5th-grade tasks to 29 5th-grade students. Both samples consisted primarily of students randomly selected from classrooms in elementary schools located in two large, urban districts.

A few of the students were children of employees at the institution where the research took place. Students varied by ethnicity, socioeconomic status, and previous experience with hands-on science. Students who completed tasks from a common shell performed them in a counterbalanced order, generally separated by two or three days. Of the 29 5th-graders, 10 completed both tasks, 9 completed Friction only, and 10 took Incline only. Of the 20 6th-graders, 17 completed both, 2 Lever only and 1 Pendulum only. Upon completion of each task, students were asked to explain the reasoning that they used on certain items. They also responded to a series of interview questions that focused on which aspects of the tasks they did and did not enjoy, what they found interesting or confusing, and whether they perceived the tasks to be good tests of their scientific knowledge. The specific questions asked are described more fully in the Results section.

Scoring rubrics were developed for each separate item in each task. For the purposes of this study, however, students were simply divided into high and low groups on the basis of their performance with respect to the central requirement of each task (e.g., designing an experiment or creating a valid cross-classification) rather than their performance on several separate items. The formation of these groups is discussed in the next section.

Results

In reporting the results of both studies, I will focus on information that relates to students' attitudes toward the tasks and how these attitudes relate to performance. These findings are preliminary. Neither tasks nor students are representative of the respective populations; therefore no inference can be made regarding generalization to other testing situations. In addition, the presence of the interviewer inevitably influences the responses that students offer. Nonetheless, these interviews do provide some potentially useful information concerning students' reactions to different test formats, and suggest areas for further research.

Because the 5th- and 6th-grade portions of the study involved different tasks and different samples of students, results are reported

separately by grade. Emphasis is on the extent to which students enjoyed each task, their feelings regarding the quality of the tasks as tests of their knowledge, and the perceived difficulty of each task.

5th-Grade Tasks

Most of the 5th-grade students enjoyed the hands-on tasks. Of the 29 students who were interviewed, 23 said the tasks were fun. All students, including the six who called the tasks boring, said that they were more interesting than the tasks they usually took in school. Reasons for calling the tasks fun included "putting things together" (n=8), using the equipment to answer the test questions (n=6), being surprised at the results (n=5), and "playing with" the equipment (n=4). Eighteen of the students said that they learned something from the activities; most of these said that they learned new concepts to which they had not previously been introduced, such as "friction," or that they learned something because they got to see what happened when they manipulated the equipment in a certain way. One student, for example, said, "I like these because you get to know the stuff better. You understand it better because you get to see it happen, not just have it told to you." Many students said that they enjoyed those aspects of the tasks that involved movement, such as watching the weight pull the truck up the inclined plane. As stated earlier, five students said that the tasks were fun because the results were surprising or unexpected: "I thought it would take the same amount of weight on each board, so it was neat to see it was different for all of them." Students who discussed this element of surprise all stated that they learned something from the tasks, which is a goal of many developers of performance assessments. Students' comments illustrate some specific features of tasks, such as movement or surprise, that can enhance student engagement. Further support for the power of these tests to engage students was acquired from responses to the question, "Did this feel like a test?" Eight students said that it felt more like a "project" or an "activity" than a test, and six more said that only the written part felt like a test. Comments indicate that the aspects of the tasks that make them fun also make them seem uncharacteristic of tests.

Students' perceptions of what made the task interesting appeared to be related to their performance. All four students who enjoyed the activities because they could "play with" the equipment, and four of the eight who said they liked them because they involved putting things together, either designed experiments that were irrelevant to the assigned task or failed to design any experiment. Many of these students spent a good deal of their time manipulating equipment without attending to the task instructions, either attempting to figure out where each part belonged, or complaining that they did not understand what to do. By contrast, of the nine students who said they enjoyed the activities because they could use the equipment to find a solution or because there was an element of surprise at the results, seven conducted appropriate experiments. Students who perform well on tasks such as these are apparently able to discover those task features that enhance their understanding of the relevant concepts and to use the equipment in the manner intended by the task developer.

All students were asked whether the tasks were "good tests of your science knowledge." For the few students who said that they did not understand the question, it was rephrased: "Do you think doing this activity allows you to show what you really know in science?" Seventeen students said the tasks were good tests of their science knowledge. Their reasons varied, but the majority believed that the hands-on nature of the tasks made them good tests, either because they involved activities similar to those in which scientists engaged, such as experimentation, or because they did not rely solely on written answers. Other students said that the fact that memorization and "cramming" were not involved made the tasks good tests, and that such tests would enable students who may not perform well on most science tests to show what they could do. One student said, "This is a good test because sometimes you think you don't know science or you're no good at it, but then you do something like this, where you have to show what you know, and then you can show you can do it." Another stated that "It's good because someone might think they don't know science cause they can't remember all the stuff they learned in school. But they could probably do this even if they don't know stuff for school. So then they'd know they

could really do science, and they'd have more confidence." Clearly, students such as these view the performance tasks as measuring something fundamentally different from traditional science tests.

Twelve students said that they did not believe the hands-on tasks to be good tests of their science knowledge. Three, who were especially frustrated with the quantity of reading and writing involved, said that they tested English rather than science. A few others said that the activities were different from what usually occurred in their science classes; for example, "No, cause I haven't really done much of this stuff in science, so I didn't have anything to think about. It didn't test all the stuff I've learned." Two said the tasks reflected mathematical knowledge moreso than science: "It's not really a science test cause it's graphing and counting and stuff. It was things I learned in math class. Oh, and it's English too, cause you gotta write sentences and make sure it's clear." Many students believed science tests should require them to recall facts and concepts, even though other students cited the lack of this requirement as a reason for describing the tasks as good tests. Evidence that some students believe science tests should require recall was supplied by two students who stated that the tasks were too easy because they allowed students to use the results of an experiment to answer the questions: "This was too easy cause it gives away the answer. You don't have to actually know anything. You can just do the experiment and find out what to put. It doesn't really show all the science I know." Both students who expressed this view stated that they liked science and normally did well on traditional science tests. Both also performed well on the hands-on tasks. Students' criteria for judging the quality of test stem, in part, from their previous experiences with testing, and would probably change if they acquired more experience with hands-on modes of assessment. In addition, perceptions of what school-acquired knowledge was applicable were related to perceived quality of the tasks as science tests: Most students who thought they were good tests reported using knowledge acquired in science class, and most others said that they used math or English more than science. It appears that the relevance of a

test to what is learned in class influences students' judgments of its quality.

Students were asked to identify the difficult and easy aspects of the tasks, and were asked, "Do you think you got the right answers?" Most students said that nothing was difficult about the hands-on part of the task; only a few mentioned difficulties identifying materials on Friction. Over two thirds said that the written part of the task was difficult, and the requirement to record steps to the experiment was viewed as the most difficult aspect. Although quality of performance varied greatly, and only three students recorded their experimental steps accurately, 11 said that they thought their answers were correct. Answers to this question seem to be unrelated to performance: Those students who performed extremely well answered yes, saying that they knew the material well. In addition, many students whose approach to the task was incorrect also reported confidence in their answers, providing reasons such as "I tested it out with the equipment, so it must be right," or "I followed the instructions." Three students claimed that there were no right answers and that any experiment they designed must be acceptable: "It's just really your opinion, how you think the experiment should go, so it can't be wrong." Perhaps these perceptions result from an absence of this kind of activity on tests that students take in school. Overconfidence in one's own responses to test items is a common phenomenon (see, for example, Fischoff, Slovic, & Lichtenstein, 1977), but may be exacerbated by perceptions of hands-on tasks as fundamentally different from traditional tests; that is, as non-testlike. Repeated exposure to this form of testing may not only alter students' perceptions of test quality and difficulty, but also their impressions concerning the nature of science instruction.

6th-Grade Tasks

Most of the 6th-graders took four tasks, two from each of two task shells. Students' reactions to each shell differed substantially, especially with regard to feelings of frustration and enjoyment. All 20 students said that they thought Lever and Pendulum, the tasks involving experiments, were fun, and that these tasks were more interesting than

the tests they usually took in school. Pendulum was the favorite of most students; 6 reported being surprised at the results, which they say made the activity fun, and several others said they enjoyed watching the pendulums swing and seeing what would happen with different strings. Most students stated that these tasks were more interesting than the tests that they usually took in school because they could use equipment and because they did not have to recall information: "You don't have to think as hard, or study, or memorize lots of stuff like on most tests." Many comments were similar to those expressed by the 5th-graders. The 6th-grade comments, however, tended to be more positive, and no 6th-grader reported boredom or frustration with the inference tasks. Although it is difficult to make inferences about why these differences exist, especially given that different samples of students provided responses for the two sets of tasks, students' remarks do suggest that the step-by-step instructions on the 6th-grade tasks resulted in less frustration than the open-ended nature of Friction and Incline. Interviews with seven 5th-graders who took Pendulum along with Friction and Incline support this; all seven said that Pendulum was less confusing and more fun because it "told me what to do" or because "the directions were easier to understand." Although open-ended tasks may be more representative of scientific activity, many students find them confusing or frustrating. This might change with further experience with such activities.

As with the 5th-grade tasks, students' comments about what made the hands-on tasks interesting were related to their performance. Students who said they enjoyed being able to use the equipment performed better than those who liked "watching what happens" or "playing with the stuff." For example, the eight students in the former group conducted the experiment correctly and completed their data sheets accurately, whereas the six who focused on playing with the equipment conducted inaccurate measurements and engaged in a good deal of activity that was irrelevant to the task at hand.

Only six students said that Pendulum and Lever felt like tests. These students emphasized the inclusion of written questions and the amount of reading required. Again, most students called the tasks

"activities" or "projects," and many considered the hands-on part and the written part to be separate components, with the written part being perceived as much less interesting. It is worth noting that several of the 5th- and 6th-grade students expressed confusion at first when they were given the equipment and the test booklet. Even after they were told that they could use the equipment to help them answer the questions in the booklet, many students asked questions such as, "Which part is the test, this (booklet) or this (equipment)?" or "How can I do these both at the same time?" The students who posed these questions reported having little experience with hands-on science activities and no experience with hands-on tests. Nonetheless, all students reported liking these activities better than traditional tests.

Reactions to the two Classification tasks tended to be less favorable. Seven of the 20 students called them "boring," and 4 of these students said they were less interesting than the tests they usually took in school. Even students who called the activities "fun" said that they were not as much fun as the Inference tasks. Reasons included: Too much reading, lack of equipment to manipulate (even though students were given objects to sort), and lack of understanding of the instructions or the purpose of the task.

Many students also asserted that the tasks were not good tests of their science knowledge. Of the 20 students taking these tasks, eight created a valid cross-classification on one or both. Three of these eight believed the activity reflected mathematics more than science, and said that they used skills acquired in math class to produce their responses. In contrast, five students who were unable to form a cross-classification said the tasks were not good tests because they were long and boring or because they did not encompass a broad range of scientific content: "This has nature stuff, but that's really only part of science. Science is electricity and experiments and stuff, not just this." Four others said that they learned most of what was needed to perform the tasks outside of school, and that the tasks were therefore inadequate science tests. Eight students said that the tasks were good tests because they included content learned in school and because they did not require much memorization or writing. Students tended to prefer

Animals to Materials, primarily because they found animals to be a more interesting subject of study and because they had been exposed to facts about animals both in and outside of school. In contrast to the other 5th- and 6th-grade tasks, three fourths of the students said that they used information learned outside of school, and many of these said this made the activities more interesting (however, many of these same students said this fact made them poor tests of their science knowledge, indicating a perception of scientific knowledge as something learned in the classroom). Students' favorite parts were choosing their own groups and finding out what the "surprise" object was.

Impressions of what made the tasks difficult or easy were related to the quality of responses. Six of the students who did not form a cross-classification said that the requirement to choose their own groups made the tasks easy; these students all formed four separate groups that represented four levels of a single variable, rather than cross-classifying on the basis of two variables. By contrast, seven of the eight who performed well said that choosing their own groups was "challenging" or "hard." Fifteen students, including six of the eight successful students, said that they thought they got the right answers. Most attributed this confidence to their knowledge about animals and nature, or to the fact that there were no single correct answers. Eleven students said the tasks felt like tests; primary reasons were that they were "boring," "hard," or "too long." Table 2 summarizes students' responses to three of the interview questions for each set of tasks.

Table 2

Summary of Responses to Science Interviews

	FRICTION/INCLINE		PENDULUM/LEVER		CLASSIFICATION	
	YES	NO	YES	NO	YES	NO
Fun?	23	6	20	0	13	7
Good test?	17	12	15	5	8	12
Feel like a test?	21	8	6	14	11	9

One additional set of observations was made with respect to the classification tasks. These tasks require students to complete a tutorial section that introduces them to cross-classification. It is important, therefore, to investigate the extent to which students found that section useful and applicable. Students were asked, "Was doing the activity with the people helpful to you when you did the part with the animals (materials)?" Of the 20 students, 12 said that it helped them. This number included the eight successful students, most of whom believed the tutorial defined the task for them or provided examples of how to approach it. The less successful students who said that it was helpful focused on the introduction of vocabulary terms, such as "group" and "property." Most of the students who did not find it helpful said that it was different from the main task, irrelevant, or too easy. For example, one student said it was irrelevant because it involved people rather than animals, and another said that it was too easy and did not adequately prepare her for the main classification task. Many of the unsuccessful students viewed the tutorial as a separate component rather than attempting to apply it to the animals or materials tasks. Typically, tests do not require students to learn something in one section and then to use this new knowledge on another section. Thus, these reactions are sensible in light of most students' experiences with test-taking.

Study 2: High School Mathematics Tests

Instruments

In a separate study, high school students took mathematics achievement items in three formats: multiple-choice, short-answer constructed-response ("open-ended") and an "extended" problem format. The multiple-choice items were those that appeared on the NELS:88 10th-grade Mathematics test forms (NCES, 1988). Three forms, each containing 40 items, were administered in the NELS:88 study. For the present study, all of the items from all three forms were placed on a single form. Because there was extensive overlap among the forms, the total number of items was 58. Most of the items focused on arithmetic and algebra, with some geometry and statistics items included as well. The

constructed-response form consisted of 17 of these items, with the response options removed. These items were selected to represent the various content domains included on the NELS:88 forms. They were also selected on the basis of their amenability to revision into the constructed-response format (e.g., items in which most of the question was contained in the response options rather than in the stem were excluded). On the constructed-response items, students were asked to supply a response and to explain their reasoning. This requirement increased time needed to respond to each item, and therefore the multiple-choice and open-ended forms took approximately the same amount of time to administer. Some of the Extended problems were selected from sets of released items for statewide testing programs, and others were created specifically for this study. They posed problems that required students to provide more extensive responses than those required by the Open-Ended problems. For example, students were asked to construct figures that had a certain area, or to explain the flaws in another student's mathematical reasoning. Like the other items, these problems involved arithmetic, algebra, and geometry.

Procedures

Twenty-nine student volunteers from a local high school participated in this study. No claim about the representativeness of this sample is made here. In fact, most of the students appeared to be high achievers, based upon their scores on the nationally-normed NELS:88 test, although they did vary in terms of level of math proficiency and reported enjoyment of mathematics as a school subject. Twenty-one students were taking Geometry at the time; five were taking Algebra II and one Algebra I. The sample included 17 females and 12 males. Students completed each test form as part of a larger study of students' mathematical reasoning. The multiple-choice and open-ended forms were administered in a counterbalanced order, and the extended problems were administered last. After taking each test, students were asked to discuss those aspects of the test that they enjoyed or found difficult, stressful, or confusing. At the end of the interview, they were asked to make comparisons among the three forms. The specific questions asked

are elaborated in the Results section. Three interviewers collected the data using a standard interview form.

Results

Unlike the 5th- and 6th-grade students, the high school students in this study were able to make direct comparisons of different testing formats because they took three versions of a test. After completing the multiple-choice (MC) and open-ended (OE) versions, students participated in interviews in which they compared these two formats. Of the 29 students who were interviewed, 20 preferred the MC test to the OE version. Although MC contained more than three times as many items as OE, 15 of these 20 students said that MC was easier. Almost all explanations for these responses referred either to the requirement that students explain their answers on OE or to the susceptibility of MC to the application of strategies such as eliminating responses, working backwards, or double checking answers by computing them and then comparing them to the available response options. Several students also recognized that the probability of answering an item correctly when they did not know the answer was much greater for MC than OE, and that partial knowledge was sometimes enough to reveal the correct answer when several response options were provided. The following statement is typical of students who preferred MC: "If I didn't know the answer, I could guess. I could make an educated guess. And I'd have a better chance of getting the problem right. And then, if I didn't know the answer, I could work backward from the answers that looked reasonable." Two students also said they liked MC because the answers were "either right or wrong," providing evidence that performance standards for multiple-choice tests are clearer to students than are those for other formats.

Of the nine students who preferred the open-ended test, five reported that they liked the greater challenge it presented. For example, one student said, "I like doing it myself, not just being given the answers," and another liked OE because it did not permit "short cuts." Two students said that they found the MC format confusing: "(I prefer) the open-ended problems, because I always get screwed up with

multiple-choice. I get confused. It's easier for me just to do my own work and come up with my own answer." One student said it was especially confusing when the answer he calculated was not among the response options. One said he found MC "tedious," and another liked OE because it allowed her to explain her answers and to show what she knew. Regardless format students prefer or which they find easier, nearly all express an awareness of the variety of strategies that can be applied to MC items.

Students were also asked, "Which test did a better job of allowing you to show what you know in math?" As earlier research has demonstrated (e.g., Bridgeman, 1992), students do not always prefer the test that they believe to be the fairer measure of what they know. The results of these interviews are consistent with results of prior studies. Although the majority of students preferred MC, only three students said that it was a better indicator of what they knew. One student said that the greater number of items on MC gave a better picture of the range of content he had covered in school, and the other two said that it enabled them to avoid "stupid mistakes" that could affect their scores on the OE test. Eleven students said the two were approximately equivalent in this regard, and fifteen thought that OE did a better job of allowing them to demonstrate their knowledge. Six said this was because OE did not allow guessing; for example, "Probably open-ended, because of the guessing thing with the multiple-choice. And this you really have to think through, and that tests your own knowledge." When asked why MC did not allow him to show what he knew, this student responded, "Well, just because of the guessing thing. It's like you can guess the answer, and you might get it right. While this (OE), you either get it right or get it wrong." Five students said that the opportunity to explain their answers made OE a better test, and three stated simply that they had to "think more" on OE. Some students noted that it was possible to award partial credit on OE, and that students with partial knowledge were therefore penalized less. Of the fifteen who said OE was the better measure of what they knew, eight reported that they preferred MC to OE. Thus, even though the majority of students named at least one way in which the OE format was a more

adequate measurement instrument, these judgments apparently did not influence their preferences.

The MC version included two item types. The "regular" items provided four or five response options from which students selected the best answer to a question. The "comparison" items required students to calculate quantities in two columns and to choose one of four responses: (A) the quantity in the first column is larger; (B) the quantity in the second column is larger; (C) the quantities are equal; and (D) there is not enough information provided to determine which is larger. These responses are provided at the beginning of the comparison section; for each item, students see only the letters A through D. The majority of students (18) said that they preferred the "regular" items, and their reasons again tended to reflect either a desire for feedback or a preference for familiar formats. Nine students said that the regular items gave them more information about whether their answers were correct, by allowing them to eliminate response options or to test their chosen response against the item stem. Eight preferred the regular items because they were familiar. One student, for example, said that she had to "keep looking back at the instructions" for the comparison items, and another complained that she had trouble remembering what A and B stood for. Three students preferred neither the regular nor the comparison items, and eight liked the comparison items better. Three of these students said that the comparison items did not require them to calculate exact answers, and three others liked the fact that there were not as many numbers to consider for any given item. One said that the comparison items were more challenging because they did not permit students to select an answer. Once again, the application of strategies such as response elimination and estimation strongly influences students' preferences for item formats.

After taking a third test, containing the "Extended" items, students responded to interview questions similar to those described above, comparing all three forms. Because one student left the interview early, results are based on interviews with 28 students. Table 3 summarizes the results of this second round of interview questions, along with responses to the first set of questions.

Table 3

Summary of Responses to Mathematics Interviews

1ST INTERVIEW	MULTIPLE-CHOICE	OPEN-ENDED	EXTENDED
Like best?	20	9	
Most difficult?	9	20	
Best test?	3	15	
2ND INTERVIEW			
Like best?	6	3	18
Most difficult?	3	10	14
Easiest?	19	2	6
Best test?	2	12	6

The majority of students liked the extended problems better than the other formats. Their reasons included the fact that they were different from tests students normally took (n=3); they required skills other than math, such as "creativity," drawing, and writing (n=7); they involved real-life scenarios (n=3); and they were more challenging (n=3) or fun (n=2). Students who did not enjoy these problems named some of these same features as perceived deficits, especially the requirement to apply other skills. Six of these ten students reported that they had no experience with items of this type, whereas only three of the students who enjoyed the Extended items reported this lack of experience.

Consistent with results from the earlier interview, students did not always name the preferred test as the one that best allowed them to demonstrate their mathematical knowledge. Only six students said Extended did this best, four of whom said it required more thinking (e.g., "you had to use your mind more"). Two students chose MC, and twelve selected OE (the other eight said the tests did not differ in this respect). The students who chose MC or OE had chosen the same format in the earlier interview, and most of their responses were consistent with their earlier explanations. In addition, several of the students who chose OE named particular features of Extended that made those problems less adequate measures of their mathematical knowledge. These features included the small number of items and the fact that

abilities other than those related to math were perceived to be required by the Extended items: "These aren't like too much from math, and it shouldn't be in paragraph form and stuff because this is almost like writing it in words like in English class." Several students said that a math test should not require other abilities, such as "creativity." Like several of the younger students who took the hands-on science tests, many of these students believe that a test that purports to measure understanding in a given subject area should not draw on skills other than those taught in that class.

Students were asked to name the format that was the easiest, and the one that was the most difficult. Once again, most students thought that MC was easiest (n=19); two named OE and six selected Extended (the remaining student said the tests were approximately equal in difficulty). The students who chose the Extended problems thought they were easiest either because they involved content other than math or because they permitted more than one correct answer. At the same time, 14 students named Extended as the most difficult; 10 chose OE and 3 MC. Most of the students who found the Extended and open-ended problems difficult emphasized the writing requirements and, on Extended, the requirement to apply knowledge other than that relating specifically to mathematics.

Because most of the students in this sample performed extremely well on both the multiple-choice and the open-ended formats, differences in performance are not examined here. The quality of responses provided on the extended problems did vary to some extent, and appeared to be related to students' perceptions of the difficulty of these items. Many of the students who found the extended problems easy failed to interpret the instructions correctly and, consequently, provided responses that were insufficient. For example, one item asked students to construct three shapes that have the same area as a given shape. A few students simply drew lines in the given shape so that it was divided into three parts, and then commented on the simplicity of the item. These students did well on the other formats, but did not always interpret the instructions on the extended problems in the way the items' creators intended. Once again, prior experience with non-traditional test

formats probably influenced the quality of responses students supplied and the approaches they took to solving each item.

Discussion

The results of both studies indicate a great deal of variability in the affective responses of students to novel assessment formats. They also suggest some possible influences on these responses. As stated earlier, this study was not intended to produce results that could be generalized to other tasks or to other samples of students, but to identify questions that might be addressed by future studies and to suggest possible hypotheses. Some of these will be discussed in this section.

First, it is essential that we determine the extent to which non-traditional assessment formats accomplish the goal of engaging students. Although this is typically not the primary objective of a test developer, it has been cited frequently as a benefit of performance assessments. We might also expect students to perform better on tests that engage them than on those they find uninteresting. Many of the students in both the math and the science studies said that the performance assessments were more enjoyable than traditional tests. Furthermore, most expressed reasons other than simply the novelty of the tests, which indicates that students' positive reactions might persist once students became accustomed to these assessments. An important finding, however, is that in many cases the novelty of the test resulted in negative reactions. This is especially true for the math study, on which several students stated that they enjoyed and felt more confident on the multiple-choice test because they were familiar with the format and with the standards for judging the adequacy of responses. Perhaps older students are more immersed in the "test culture" that exists in schools, and are more aware of the high stakes attached to many test performances. In any case, students' perceptions of how enjoyable and difficult a test is are likely to change as they acquire experience with new formats. The effect of novelty on students' attitudes toward performance assessments may be investigated as such assessments are gradually introduced into classrooms and schools.

Students' responses also indicate that the nature of the engagement is important to consider. Most teachers and task developers would no doubt prefer that students enjoy an activity because of its power to enhance learning rather than because it involves objects that are "fun to play with." In the science study, although most of the students found the hands-on tasks to be enjoyable, their reasons revealed different levels of attention to the learning and discovery aspects of the tasks. Students who view an activity as an opportunity to play are less likely to focus their efforts on producing quality responses, and Study 1 suggests that their performance will generally be poorer. In contrast, those who are attuned to the opportunity for scientific exploration provided by a set of equipment will most likely perform better. Once again, experience with this type of test format will probably influence the ways in which students respond.

Study 1 also revealed that engagement often results from a focus on process rather than on content. Most students clearly find activities such as memorizing tedious, and welcome the opportunity to perform an activity that places minimal emphasis on knowledge of facts. It is neither possible nor desirable, however, for assessments to be "content-free." The tasks used in this study were designed to be performed by students with widely varying school experiences. As assessments are built to more closely match curriculum, it is likely that they will place greater demands on specific content knowledge. Will a hands-on activity that calls on factual knowledge be less engaging than one that makes no such demands? This question is important to address as performance assessments begin to be used more frequently and as tests demand the application of a greater level of content knowledge.

Students' perceptions of the validity and fairness of various item formats must also be addressed as paper-and-pencil tests are replaced by other item types. In both the math and the science studies, several students expressed reservations about whether the performance assessments were really measuring what they purported to measure. Students, like teachers and test developers, are aware that a test taker should not be penalized on a test of one kind of achievement for lack of

experience in an unrelated area. At the same time, it is recognized that most scientific and mathematical activity that takes place outside the classroom requires the integration of many skills, not all of which are explicitly taught in math or science courses. If the use of performance assessments is accompanied by changes in curriculum that include a greater level of integration among the various school subjects, students will probably be less likely to perceive such tests as unfair. In any case, perceptions of fairness and validity are important areas for future investigation. As performance assessments are developed and implemented on a large scale, valuable information will be obtained by eliciting reactions of both students and teachers, and identifying ways in which these reactions enhance or compromise the validity of the assessment.

References

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. Journal of Educational Measurement, 29, 253-271.

Check, J. F. (1982). Relative merits of test items as perceived by college students. College Student Journal, 16, 100-104.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. Review of Educational Research, 58, 438-481.

D'Ydewalle, G., Swerts, A., & De Corte, E. (1983). Study time and test performance as a function of test expectations. Contemporary Educational Psychology, 8, 55-67.

Fischoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. Journal of Experimental Psychology: Human Perception and Performance, 3, 552-564.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18 (9), 27-32.

Frederiksen, N. (1984). The real test bias. American Psychologist, 39, 193-202.

Gardner, H. (1992). Assessment in context: The alternative to standardized tests. In B. R. Gifford & M. C. O'Connor (Eds.), Changing assessments: Alternative views of aptitude, achievement and instruction (pp. 78-119). Boston: Kluwer.

Koretz, D., Lewis, E., Skewes-Cox, T., & Burstein, L. (1992). Omitted and not-reached items in mathematics in the 1990 National Assessment of Educational Progress (CSE Technical Report No. 357). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20 (8), 15-21.

Lundeberg, M. A., & Fox, P. W. (1991). Do laboratory findings on test expectancy generalize to classroom outcomes? Review of Educational Research, 61, 94-106.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New Tools for Educational Reform. In B. R. Gifford & M. C. O'Connor (Eds.), Changing assessments: Alternative views of aptitude, achievement and instruction (pp. 37-75). Boston: Kluwer.

Rocklin, T. (1992). A multidimensional scaling study of college students' perceptions of test item formats. Applied Measurement in Education, 5, 123-136.

Snow, R. E. (1989). Toward assessment of cognitive and conative structures in learning. Educational Researcher, 18 (9), 8-14.

Snyder, B. R. (1971). The hidden curriculum. Cambridge, MA: MIT Press.

Wiggins, G. (1992). Creating tests worth taking. Educational Leadership, 49 (8), 26-33.